

Examples of inconsistency in optimization by expected improvement

D. Yarotsky^{*†}

March 31, 2012

Abstract

We consider the 1D Expected Improvement optimization based on Gaussian processes having spectral densities converging to zero faster than exponentially. We give examples of problems where the optimization trajectory is not dense in the design space. In particular, we prove that for Gaussian kernels there exist smooth objective functions for which the optimization does not converge on the optimum.

Contents

1	Introduction	1
2	Results	5
3	A numerical example	8
4	Proof of Theorem 1	9
5	Proof of Theorem 2	11
6	Proof of Theorem 3	16

1 Introduction

The optimization problem. Consider a global “black-box” optimization problem

$$f(x) \longrightarrow \min_{x \in \mathcal{D}} \quad (1)$$

^{*}Datadvance llc, Moscow, yarotsky@datadvance.net

[†]Institute for Information Transmission Problems, Moscow, yarotsky@iitp.ru

For the moment, suppose that \mathcal{D} is a compact metric space, and f a continuous real-valued function on \mathcal{D} , so that the minimum $f^* = \min_{x \in \mathcal{D}} f(x)$ exists. Consider an optimization procedure seeking this minimum. In black-box optimization, such a procedure consists of a sequence of iterations; each iteration suggests for evaluation a new point of the set \mathcal{D} based on the already observed values of the objective function f . More precisely, we can say that, algorithmically, optimization is defined by the initial point $x_1 \in \mathcal{D}$ and a family of mappings

$$\mathcal{A}_K : (\mathcal{D} \times \mathbb{R})^K \rightarrow \mathcal{D}, \quad K = 1, 2, \dots$$

The optimization trajectory $\{x_K\}_{K=1}^\infty$ is then determined by relations

$$x_{K+1} = \mathcal{A}_K\left(\{x_k, f(x_k)\}_{k=1}^K\right), \quad K = 1, 2, \dots \quad (2)$$

Any practical optimization is terminated at some step K , and the approximate minimum f_K^* is then defined by

$$f_K^* := \min_{k=1, \dots, K} f(x_k).$$

It is then natural to call optimization *consistent* if

$$\lim_{K \rightarrow \infty} f_K^* = f^*.$$

The following proposition is a very simple but important criterion of consistency on the space of continuous functions [16].

Proposition 1. *An optimization algorithm defined by mappings \mathcal{A}_K is consistent for all $f \in C(\mathcal{D})$ if and only if for any continuous f the trajectory $\{x_K\}_{K=1}^\infty$ generated by (2) is dense in \mathcal{D} .*

The sufficiency is clear; the necessity follows since any continuous function can be modified, preserving its continuity, in any open set so as to make the function attain its optimum in this open set.

In many practical applications, the objective function f is expensive to evaluate, and the mappings \mathcal{A} can then be quite complex and resource-intensive; in particular they often involve solving auxiliary optimization problems. A popular modern approach to global black-box optimization is stochastic Bayesian optimization where these auxiliary problems are stated using some prior assumptions of probabilistic nature. In this paper we will consider one of the most natural and well-known methods of this type – optimization by Expected Improvement [4–6, 10, 11, 14, 15]

The Expected Improvement algorithm (EI). In this method, we think of the optimized function f as a realization of a stochastic process $(\xi_x)_{x \in \mathcal{D}}$. Assuming the probability measure associated with the process is known, we define the mappings \mathcal{A}_K by maximizing the expectation of the improvement in the best known value of the objective function resulting from its additional evaluation, conditioned on the set $\{\xi_{x_k} = f(x_k)\}_{k=1}^K$. Precisely, we define

$$\mathcal{A}_K\left(\{x_k, f(x_k)\}_{k=1}^K\right) = \arg \max_{x \in \mathcal{D}} I_{K; \{x_k, f(x_k)\}_{k=1}^K}(x), \quad (3)$$

where

$$I_{K;\{x_k, f(x_k)\}_{k=1}^K}(x) = \mathbb{E}\left(f_K^* - \min(f_K^*, \xi_x) \mid \{\xi_{x_k} = f(x_k)\}_{k=1}^K\right).$$

In practice, the stochastic process ξ_x is usually Gaussian, which allows one to numerically solve the optimization problem

$$I_{K;\{x_k, f(x_k)\}_{k=1}^K}(x) \longrightarrow \max_{x \in \mathcal{D}} \quad (4)$$

for moderate values of K . Namely, assume that ξ_x is a centered Gaussian process with the covariance

$$G(x, y) = \mathbb{E}(\xi_x \xi_y).$$

Then ξ_x conditioned on $\{\xi_{x_k} = f(x_k)\}_{k=1}^K$ is also a Gaussian random variable:

$$\xi_x \mid \{\xi_{x_k} = f(x_k)\}_{k=1}^K \sim \mathcal{N}(m_{x;\{x_k, f(x_k)\}_{k=1}^K}, \sigma_{x;\{x_k\}_{k=1}^K}^2),$$

where m and σ^2 denote the conditional mean and variance. Note that since the process is Gaussian, the variance depends on $\{x_k\}_{k=1}^K$ but not on $\{f(x_k)\}_{k=1}^K$. A straightforward calculation shows that

$$m_{x;\{x_k, f(x_k)\}_{k=1}^K} = \mathbf{g}_{K,x}^t \mathbf{G}_K^{-1} \mathbf{f}_K, \quad (5)$$

$$\sigma_{x;\{x_k\}_{k=1}^K}^2 = G(x, x) - \mathbf{g}_{K,x}^t \mathbf{G}_K^{-1} \mathbf{g}_{K,x}, \quad (6)$$

where

$$\begin{aligned} \mathbf{f}_K &= (f(x_1), \dots, f(x_K))^t, \\ \mathbf{g}_{K,x} &= (G(x, x_1), \dots, G(x, x_K))^t, \\ \mathbf{G}_K &= (G(x_k, x_l))_{k,l=1}^K. \end{aligned}$$

Throughout the paper, we will assume that the kernel $G(x, y)$ is strictly positive definite, which in particular ensures that \mathbf{G}_K in (5),(6) is invertible.

If G is continuous, then the conditional mean and variance continuously depend on x , which implies existence of the maximum in (4). The maximum can be attained at more than one point; any of them can be taken as x_{K+1} . Up to this ambiguity, the EI algorithm is completely determined by the kernel G .

Note that if the kernel G is strictly positive definite, then

$$I_{K;\{x_k, f(x_k)\}_{k=1}^K}(x) \begin{cases} = 0, & x \in \{x_k\}_{k=1}^K, \\ > 0, & x \notin \{x_k\}_{k=1}^K, \end{cases}$$

so that the maximizer $x_{K+1} \notin \{x_k\}_{k=1}^K$, i.e., all the points of the trajectory $\{x_k\}_{k=1}^\infty$ are different.

Consider the Hilbert space $L^2(\Omega, \mathbf{P})$, where (Ω, \mathbf{P}) is the probability space on which the process ξ_x is defined. Then one can geometrically interpret the conditional variance $\sigma_{x;\{x_k\}_{k=1}^K}^2$

as the squared distance between the vector ξ_x and the linear span of the vectors $\{\xi_{x_k}\}_{k=1}^K$ in $L^2(\Omega, \mathbf{P})$.

Practical implementations of the EI algorithm often use somewhat more complex modelling than described above, based on kriging [6]. This approach includes additional polynomial trends in the model; also, the covariance function is assumed to depend on a few parameters which are adjusted at each iteration using cross-validation or maximum likelihood estimates. We will not consider these complications in this paper.

We will fix a kernel G and will treat the EI algorithm described above as ideally implemented with this kernel, in the sense that the auxiliary problem (4) is assumed to be exactly solved at each iteration K . We will then be interested in the convergence properties of the resulting sequences x_K and $f(x_K)$.

Previous rigorous results. EI is a popular approach to global optimization in modern engineering applications, but not much has been proved about it rigorously. If ξ_x is the Wiener process or its stationary version, the Ornstein-Uhlenbeck process, on a segment in \mathbb{R} , then, using the Markov property, it is not hard to check that the EI optimization is consistent for continuous objective functions, see [9]. In [17, 18], Vazquez and Bect considered the general case of compact subsets of \mathbb{R}^n and proved the convergence of the EI algorithm for sufficiently “rough” stationary processes ξ_x , on objective functions f from the reproducing-kernel Hilbert space (RKHS) associated with ξ_x .

As an intermediate step in their proof, these authors consider what they call the *No-Empty-Ball* (NEB) property of the process ξ :

Definition 1. *The process ξ_x is said to have the NEB property if for all sequences $\{x_k\}_{k=1}^\infty$ (not necessarily given by 2) and all points x in \mathcal{D} the following two conditions are equivalent:*

1. *x belongs to the closure of $\{x_k\}_{k=1}^\infty$;*
2. *the conditional variance $\sigma_{x; \{x_k\}_{k=1}^K}^2 \rightarrow 0$ as $K \rightarrow \infty$.*

The first condition clearly implies the second for processes with a continuous covariance function, but the opposite direction is more subtle. In particular, Vazquez and Bect prove that the NEB property is violated by Gaussian processes with a Gaussian covariance function. They show, however, that the NEB property holds for a stationary process provided its spectral density goes to zero sufficiently slowly, namely if its inverse is polynomially bounded. Additionally, they show that if a Gaussian process has the NEB property and the objective function is from the corresponding RKHS, then the optimization trajectory is dense in \mathcal{D} , and hence optimization is consistent on this space.

Vazquez and Bect also show that for Gaussian processes with the NEB property the optimization trajectory is dense almost surely, if the optimized function is a realization of the process.

Recently, Bull [1] has obtained rigorous convergence rates for objective functions from the RKHS of the process.

Some rigorous results are also available for certain stochastic optimization algorithms different from but closely related to EI, see, e.g., Gutmann [3].

Finally, though in this article we don't consider covariance functions with adaptively adjusted parameters, we mention that these more general kinds of EI optimization are known to be inconsistent in some cases [1, 9].

2 Results

As discussed above, the existing rigorous results about convergence of the EI optimization are mostly proofs of convergence under certain assumptions, namely when the NEB property holds and/or the objective function belongs to the RKHS associated with the process. At the same time, little is known rigorously about (in)consistency of the EI optimization when these assumptions are violated, though, for example, the Gaussian kernel is one of the most common kernels used in practical modelling [2], while in engineering applications one rarely expects strong regularity of the objective function.

Vazquez and Bect [17, 18] conjecture consistency for all continuous objective functions provided the process has the NEB property. The result of Locatelli [9] confirms this in the case of the Wiener process.

The goal of this paper is to examine convergence of the EI algorithm for analytic Gaussian processes. More precisely, we will consider kernels with spectral densities which very rapidly converge to 0; this property is related to analyticity by Paley-Wiener-type theorems (see, e.g., [7], page 209). Our main result is a class of examples demonstrating some lack of consistency of the EI optimization in this case, for objective functions which are not analytic. We thus show, in particular, that the EI optimization cannot be fully consistent if both the NEB and RKHS assumptions are dropped.

We will consider only 1D models in this paper and let $\mathcal{D} = [-1, 1]$. We consider a translation invariant covariance G , i.e.

$$G(x', x'') = G(x' - x'', 0) \equiv G(x' - x''),$$

and assume that it has a spectral density $\hat{G}(t), t \in \mathbb{R}$:

$$G(x) = \int_{\mathbb{R}} \hat{G}(t) e^{itx} dt, \quad \hat{G}(t) = \frac{1}{2\pi} \int_{\mathbb{R}} G(x) e^{-itx} dx.$$

Since G is real and even, such is \hat{G} : $\hat{G}(t) = \hat{G}(-t) \in \mathbb{R}$. We assume that $\hat{G}(t) > 0$ for all t , so that the kernel G is strictly positive definite.

We start by showing, as a preparation for the main result, that if

$$\hat{G}(t) \leq c_0 e^{-c|t|}, \tag{7}$$

with some $c_0, c > 0$, then the process ξ_x does not have the NEB property. Condition (7) implies, in particular, that G is analytic in the strip $|\operatorname{Im}(z)| < c$.

Theorem 1. *Let $(\xi_x)_{x \in [-1,1]}$ be a centered stationary Gaussian process defined on a probability space (Ω, \mathbf{P}) . Suppose that the spectral density \hat{G} of the process satisfies condition (7) with some $c_0, c > 0$. Let A be any infinite subset of the segment $[-1, 1]$. Then all the random variables $(\xi_x)_{x \in [-1,1]}$ belong to the closed linear span of the random variables $(\xi_y)_{y \in A}$ in $L^2(\Omega, \mathbf{P})$.*

We next indicate a class of optimization problems where the EI optimization trajectory is provably not dense in $[-1, 1]$.

In proving this result, we have found especially useful a pair of asymptotic bounds for the conditional variance, which we will state now as a separate theorem.

Suppose that the spectral density is represented in the form

$$\hat{G}(t) = e^{-S(|t|)} = e^{-T(\ln |t|)} \quad (8)$$

with some functions $S \in C^2(\mathbb{R}_+)$, $T \in C^2(\mathbb{R})$. We will assume that

$$S'(t), S''(t) \geq 0 \quad \text{for } t > 0, \quad (9)$$

and

$$S'(t) \rightarrow +\infty \quad \text{as } t \rightarrow +\infty. \quad (10)$$

Condition (9) implies, in particular, that T is convex, since

$$T''(s) = (S(e^s))'' = e^s S'(e^s) + e^{2s} S''(e^s) \geq 0. \quad (11)$$

Also,

$$T'(s) \rightarrow +\infty \quad \text{as } s \rightarrow +\infty, \quad (12)$$

since

$$T'(s) = e^s S'(e^s). \quad (13)$$

Let T^* be the Legendre transform of T :

$$T^*(q) = \max_{s \in \mathbb{R}} (qs - T(s)).$$

Then, by (12), $T^*(q)$ is finite for all sufficiently large q ; the point s^* where the maximum is attained satisfies the condition $T'(s^*) = q$.

Theorem 2. *Suppose that the spectral density of the covariance function G is represented in the form (8) so that conditions (9), (10) hold. Then, for sufficiently large K , the following inequalities hold for any $K + 1$ different points $x, x_1, \dots, x_K \in [-1, 1]$:*

$$e^{-K} \leq \frac{\sigma_{x; \{x_k\}_{k=1}^K}^2}{e^{F(K)} \prod_{k=1}^K |x - x_k|^2} \leq e^{2K}, \quad (14)$$

where

$$F(K) = T^*(2K + 1) - (2K + 1) \ln K.$$

Furthermore, $F(K)$ monotonically decreases for sufficiently large K , and

$$\frac{F(K)}{K} \rightarrow -\infty \quad \text{as } K \rightarrow +\infty. \quad (15)$$

An example of a family of spectral densities covered by this theorem is

$$\widehat{G}(t) = e^{-a|t|^b} \quad (16)$$

with $a > 0, b > 1$. In particular, with $b = 2$ this gives the Gaussian covariance functions

$$G(x) = \frac{1}{2\sqrt{\pi a}} e^{-\frac{x^2}{4a}}. \quad (17)$$

Spectral densities (16) correspond to

$$\begin{aligned} S(|t|) &= a|t|^b, \\ T(s) &= ae^{bs}, \end{aligned}$$

so that conditions (9),(10) hold for all $a > 0, b > 1$. We find in this case

$$\begin{aligned} T^*(q) &= \frac{q}{b} \left(\ln \frac{q}{ab} - 1 \right), \\ F(K) &= \frac{2K+1}{b} \left(\ln \frac{2K+1}{ab} - b \ln K - 1 \right). \end{aligned}$$

We state now our main result on the EI optimization. Recall that if G is a positive definite kernel, then $G(0) = \max_{x \in \mathbb{R}} G(x)$.

Theorem 3. *Under assumptions of Theorem 2, consider optimization problem (1) on $\mathcal{D} = [-1, 1]$ with the objective function*

$$f = -G.$$

Suppose that the EI optimization with the kernel G starts from the point $x_1 = 0$ (i.e., the point where the minimum is already attained). Then the optimization trajectory $\{x_k\}_{k=1}^\infty$ converges to 0; in particular the trajectory is not dense in $[-1, 1]$. Moreover, for sufficiently large K

$$e^{2^K F(K)} \leq |x_{K+1}| \leq e^{F(K)/3}, \quad (18)$$

where F is as defined in Theorem 2.

As pointed out in Proposition 1, Theorem 3 implies that there are continuous objective functions for which the EI optimization is inconsistent. Such functions can be obtained by modifying the objective function $-G$ on a set not containing points of the corresponding trajectory $\{x_k\}_{k=1}^\infty$. Recall that under assumptions of Theorems 2 and 3 the kernel G is analytic. We cannot modify $-G$ preserving its analyticity, but can modify it preserving its infinite smoothness. Recalling the family of examples discussed above, we obtain, in particular, the following corollary regarding the Gaussian covariance function.

Corollary 1. *For any Gaussian covariance function (17), there exists an objective function $f \in C^\infty([-1, 1])$ such that the EI optimization of f starting with $x_1 = 0$ is not consistent.*

We briefly discuss now practical implications of these results.

On the one hand, there are certain caveats to their practical interpretation. First, we consider only the simplest version of the EI optimization in 1D, while real applications are mostly higher-dimensional. Second, realistic optimization budgets may be too low in many problems for the indicated asymptotic behavior to be relevant. Third, the theoretical consistency, as such, may in principle be restored by trivial adjustments of the algorithm, e.g., by occasionally alternating the EI trajectory with a fixed dense sequence in \mathcal{D} .

Nevertheless, our results suggest that, in general, EI algorithms with analytic kernels are not reliable beyond a narrow class of very smooth functions – at least hard to justify theoretically. Moreover, they may be prone to early ill-conditioning and numerical instabilities due to excessive accumulation of trajectory points (see also the numerical example below). It appears that for practical numerical optimization of generic objective functions, if EI is to be applied, then a more reliable choice for the covariance would be a rough kernel with an inverse polynomial falloff of the spectral density, for example from the Matérn family (see, e.g., [13]).

In the next section we report a numerical test of Theorem 3. Then, in sections 4–6 we provide the proofs of Theorems 1–3.

3 A numerical example

To confirm Theorem 3, we report direct numerical results of the EI optimization of the objective function $f(x) = -e^{-x^2}$ performed with the kernel $G(x) = e^{-x^2}$.

In short, our numerical procedure is as follows. At each iteration, for any trial point x we compute the parameters m_x, σ_x^2 of the associated posterior Gaussian variable by explicitly using formulas (5),(6). We then compute the expected improvement at x using the well-known formula (see [5])

$$I_K(x) = (f_K^* - m_x)\Psi\left(\frac{f_K^* - m_x}{\sigma_x}\right) + \sigma_x\psi\left(\frac{f_K^* - m_x}{\sigma_x}\right),$$

where ψ and Ψ are the standard normal density and cumulative distribution function, respectively. To optimize $I_K(x)$ over x , we simply sample x uniformly on a logarithmic scale: precisely, we try $x = \pm e^{-l\epsilon}$, where $\epsilon = 0.02$ and $l = 0, 1, \dots, 10^4$.

We should point out that this numerical procedure is quite unstable for our kernel and objective function. As the posterior variance of the process rapidly converges to 0 and the trajectory $\{x_K\}$ to $x_1 = 0$, computation of the expected improvement involves, in several places, subtraction of almost equal quantities, in particular in (6). Also, the matrices \mathbf{G}_K quickly get ill-conditioned. As a result, precision of, for example, the usual “double” floating point format, which has the 53-bit significand (approximately 16 decimal digits), is exhausted very soon during this optimization. For this reason, we perform our test with the extended precision of 300 decimal digits, using the free library MPMATH [12] for that purpose.

The first 10 elements of the trajectory appear then to be reasonably reliably computed, and are shown in Table 1, together with the corresponding expected improvements. This

result confirms Theorem 3, also suggesting the actual asymptotic of x_K is closer to the lower rather than upper bound in (18).

K	x_K	$I_{K-1;\{x_k, f(x_k)\}_{k=1}^{K-1}}(x_K)$
1	0	—
2	-0.63	0.16
3	0.77	0.13
4	0.23	0.025
5	-0.1	0.0013
6	0.0036	3.4e-06
7	-7.3e-06	1.4e-11
8	2.8e-11	2.2e-22
9	-4.1e-22	4.5e-44
10	7.9e-44	1.7e-87

Table 1: The first 10 elements of the EI optimization trajectory with the respective expected improvements, for the kernel $G(x) = e^{-x^2}$ and the objective function $f = -G$.

4 Proof of Theorem 1

Let the Hilbert space \mathcal{H} be the closed span of the Gaussian random variables $(\xi_x)_{x \in [-1, 1]}$ in $L^2(\Omega, \mathbf{P})$. We use the canonical isometry between \mathcal{H} and $L^2(\mathbb{R}, \widehat{G})$:

$$\xi_x \in \mathcal{H} \mapsto \phi_x \in L^2(\mathbb{R}, \widehat{G}), \quad (19)$$

where

$$\phi_x(t) := e^{itx},$$

so that

$$\langle \xi_x, \xi_{x'} \rangle_{\mathcal{H}} = G(x - x') = \int_{\mathbb{R}} e^{itx} e^{-itx'} \widehat{G}(t) dt = \langle \phi_x, \phi_{x'} \rangle_{L^2(\mathbb{R}, \widehat{G})}.$$

In terms of this isometry, the claim of Theorem 1 is that for any $x \in [-1, 1]$ the function ϕ_x can be approximated in $L^2(\mathbb{R}, \widehat{G})$ by finite linear combinations of functions $(\phi_y)_{y \in A}$.

We first prove the following

Lemma 1. *Let x be any point in $[-1, 1]$, and $\{x_k\}_{k=1}^{\infty}$ any infinite sequence of points in $[-1, 1]$ such that $x_k \neq x_l$ for $k \neq l$ and $|x - x_k| < \frac{c}{4}$ for all k , where c is from (7). Then, assuming (7), ϕ_x can be approximated in $L^2(\mathbb{R}, \widehat{G})$ with arbitrary accuracy by finite linear combinations of ϕ_{x_k} .*

Proof. By the theory of polynomial interpolation (see, e.g., [8]), for any positive integer K we can choose coefficients $\lambda_{K,1}, \dots, \lambda_{K,K}$ such that for any polynomial p with $\deg p < K$ we have

$$p(x) = \sum_{k=1}^K \lambda_{K,k} p(x_k), \quad (20)$$

namely,

$$\lambda_{K,k} = \prod_{\substack{l=1 \\ l \neq k}}^K \frac{x - x_l}{x_k - x_l}.$$

The r.h.s. of (20) is a polynomial in x of degree $< K$. If a function $p(x)$ is not a polynomial of degree $< K$, then the difference between the left and right sides of (20) can be interpreted as the error of polynomial interpolation and written in terms of divided differences of p :

$$p(x) - \sum_{k=1}^K \lambda_{K,k} p(x_k) = p[x, x_1, \dots, x_K] \prod_{k=1}^K (x - x_k).$$

By the Hermite–Genocchi formula,

$$|p[x, x_1, \dots, x_K]| \leq \frac{\left\| \frac{d^K p}{dx^K} \right\|_{\text{conv}(x, x_1, \dots, x_K)}}{K!},$$

where $\|\cdot\|_{\text{conv}(x, x_1, \dots, x_K)}$ denotes the maximum over the convex hull of the points x, x_1, \dots, x_K . In particular, if $p(x) = e^{ixt}$ with some $t \in \mathbb{R}$, then

$$|p[x, x_1, \dots, x_K]| \leq \frac{t^K}{K!}.$$

Accordingly,

$$\left| e^{ixt} - \sum_{k=1}^K \lambda_{K,k} e^{ix_k t} \right| \leq \frac{t^K}{K!} \prod_{k=1}^K |x - x_k|$$

and hence

$$\left\| \phi_x - \sum_{k=1}^K \lambda_{K,k} \phi_{x_k} \right\|_{L^2(\mathbb{R}, \hat{G})}^2 \leq \int_{\mathbb{R}} \frac{t^{2K}}{(K!)^2} \hat{G}(t) dt \prod_{k=1}^K |x - x_k|^2. \quad (21)$$

Now recall that $\hat{G}(t) \leq c_0 e^{-c|t|}$ with some $c_0, c > 0$, and $|x - x_k| < \frac{c}{4}$. Then

$$\begin{aligned} \left\| \phi_x - \sum_{k=1}^K \lambda_{K,k} \phi_{x_k} \right\|_{L^2(\mathbb{R}, \hat{G})}^2 &\leq \frac{2c_0}{c^{2K+1} (K!)^2} \left(\frac{c}{4} \right)^{2K} \int_0^{+\infty} s^{2K} e^{-s} ds \\ &\leq \frac{2c_0 (2K)!}{c 4^{2K} (K!)^2} \\ &= O(2^{-2K}) \xrightarrow{K \rightarrow \infty} 0, \end{aligned}$$

where we used Stirling's formula in the last step. □

Now, let B denote the set of all those points x in $[-1, 1]$ for which ϕ_x can be approximated by linear combinations of $(\phi_y)_{y \in A}$. We prove that $B = [-1, 1]$ in several steps.

Statement 1. B has a non-empty interior.

Indeed, since $A \subset [-1, 1]$ is infinite, we can find an interval of length $\frac{\varepsilon}{4}$ in $[-1, 1]$ that contains infinitely many points of A . Then, by the above lemma, any point of this interval belongs to B .

Statement 2. If x belongs to the interior of B and $|x' - x| < \frac{\varepsilon}{4}$ for some $x' \in [-1, 1]$, then x' also belongs to the interior of B .

Indeed, it follows from the hypothesis that we can find infinitely many distinct points x_k in B such that $|x' - x_k| < \frac{\varepsilon}{4}$ for all of them. By the above lemma, $\phi_{x'}$ can then be approximated by finite linear combinations of ϕ_{x_k} . But, since $x_k \in B$, any ϕ_{x_k} can in turn be approximated by finite linear combinations of $(\phi_y)_{y \in A}$. It follows that $\phi_{x'}$ can be approximated by finite linear combinations of $(\phi_y)_{y \in A}$, i.e., $x' \in B$.

Now, in the above argument, x' could be replaced by any x'' sufficiently close to x' so that $|x'' - x| < \frac{\varepsilon}{4}$ still holds, and we would get $x'' \in B$. It follows that x' belongs not only to B , but even to the interior of B .

Statement 3. $B = [-1, 1]$.

This follows immediately from statements 1 and 2.

This completes the proof of Theorem 1.

5 Proof of Theorem 2

We use again the canonical isometry (19) to express the conditional variance $\sigma_{x; \{x_k\}_{k=1}^K}^2$ as

$$\sigma_{x; \{x_k\}_{k=1}^K}^2 = \min_{\lambda_1, \dots, \lambda_K} \int_{\mathbb{R}} \left| e^{ixt} - \sum_{k=1}^K \lambda_k e^{ix_k t} \right|^2 \widehat{G}(t) dt.$$

We start now with the proof of the upper bound in (14). We already know from the proof of Theorem 1 that (see (21))

$$\sigma_{x; \{x_k\}_{k=1}^K}^2 \leq \frac{1}{(K!)^2} \prod_{k=1}^K |x - x_k|^2 \int_{\mathbb{R}} t^{2K} \widehat{G}(t) dt. \quad (22)$$

We substitute $t = \pm e^s$ in the integral on the r.h.s.:

$$\int_{\mathbb{R}} t^{2K} \widehat{G}(t) dt = 2 \int_{\mathbb{R}} e^{(2K+1)s - T(s)} ds. \quad (23)$$

We can now derive an upper bound for this integral using a basic form of the Laplace method. Consider the function $\widetilde{T}_K(s) := (2K+1)s - T(s)$ which is concave by (11). Let

$$s_K^* = \arg \max \widetilde{T}_K(s). \quad (24)$$

Using $\tilde{T}'_K(s_K^*) = 0$ and $\tilde{T}''_K = -T''$, we can write

$$\begin{aligned}
\tilde{T}_K(s) &= \tilde{T}_K(s_K^*) + \tilde{T}'_K(s_K^*)(s - s_K^*) + \int_{s_K^*}^s \left(\int_{s_K^*}^{s_1} \tilde{T}''_K(s_2) ds_2 \right) ds_1 \\
&= \tilde{T}_K(s_K^*) - \int_{s_K^*}^s \left(\int_{s_K^*}^{s_1} T''(s_2) ds_2 \right) ds_1 \\
&\leq \tilde{T}_K(s_K^*) - \int_{s_K^*}^s \left(\int_{s_K^*}^{s_1} \chi''(s_2 - s_K^*) ds_2 \right) ds_1 \\
&= \tilde{T}_K(s_K^*) - \chi(s - s_K^*) \\
&= T^*(2K+1) - \chi(s - s_K^*), \quad \text{for all } s \in \mathbb{R},
\end{aligned} \tag{25}$$

for any C^2 function χ such that $\chi(0) = \chi'(0) = 0$ and

$$\chi''(s) \leq T''(s_K^* + s) \quad \text{for all } s. \tag{26}$$

It follows then from (25) that

$$\int_{\mathbb{R}} e^{(2K+1)s - T(s)} ds \leq c_0 e^{T^*(2K+1)},$$

where $c_0 = \int_{\mathbb{R}} e^{-\chi(s)} ds$. By (10),(11), $T''(s) \xrightarrow{s \rightarrow +\infty} +\infty$, so we can choose a χ such that $c_0 = \frac{1}{2}$ while (26) and hence (25) hold for all sufficiently large K ; for example

$$\chi(s) = c_1 \cdot \begin{cases} 6s^2 - s^4, & |s| \leq 1, \\ 8|s| - 3, & |s| > 1, \end{cases}$$

with the appropriate constant c_1 . Using Stirling's formula, we then get from (22),(23), for sufficiently large K ,

$$\sigma_{x; \{x_k\}_{k=1}^K}^2 \leq e^{T^*(2K+1) - (2K+1) \ln K + 2K} \prod_{k=1}^K |x - x_k|^2,$$

which is the upper bound in (14).

To prove the lower bound in (14), we will use the following lemma.

Lemma 2. *Let $z \in \mathbb{C}$ and $\{z_k\}_{k=1}^K \subset \mathbb{C}$. Let*

$$v = (1, z, z^2, \dots, z^K) \in \mathbb{C}^{K+1}.$$

Similarly, let

$$v_k = (1, z_k, z_k^2, \dots, z_k^K) \in \mathbb{C}^{K+1}, \quad k = 1, \dots, K.$$

Then the standard l^2 distance ρ in \mathbb{C}^{K+1} between v and the linear span of $\{v_k\}_{k=1}^K$ equals

$$\rho = \frac{\prod_{k=1}^K |z - z_k|}{\left(1 + \sum_{q=1}^K \left| \sum_{1 \leq k_1 < \dots < k_q \leq K} \prod_{t=1}^q z_{k_t} \right|^2 \right)^{1/2}}. \tag{27}$$

Proof. We have

$$\rho^2 = \frac{g(v, v_1, \dots, v_K)}{g(v_1, \dots, v_K)}, \quad (28)$$

where $g(\cdot)$ denotes the Gram determinant of the given system of vectors. Since v_k and v are $(K+1)$ -dimensional, $g(v, v_1, \dots, v_K)$ can be computed simply from the Vandermonde determinant for z, z_1, \dots, z_K :

$$g(v, v_1, \dots, v_K) = \prod_{0 \leq k < l \leq K} |z_k - z_l|^2, \quad (29)$$

where we have denoted $z_0 \equiv z$. In order to compute $g(v_1, \dots, v_K)$, we note first that it can be expressed, by the Cauchy-Binet formula, as

$$g(v_1, \dots, v_K) = \sum_{s=0}^K |\Delta_{K,s}|^2, \quad (30)$$

where $\Delta_{K,s}$ is the $K \times K$ minor of the $K \times (K+1)$ matrix $(z_k^t)_{k=1, t=0}^{K,K}$ obtained by removing the row $(z_k^s)_{k=1}^K$. Note that $\Delta_{K,K}$ is the usual Vandermonde determinant, and $\Delta_{K,0} = \Delta_{K,K} \prod_{k=1}^K z_k$. We can compute $\Delta_{K,s}$ for any s in a way similar to the usual inductive evaluation of the Vandermonde determinant. Namely, define for brevity

$$\mu_s(t) = \begin{cases} t, & t < s; \\ t+1, & t \geq s. \end{cases}$$

Then for $0 < s < K$ we have, performing linear transformations with rows and columns,

$$\begin{aligned} \Delta_{K,s} &= \det \left(z_k^{\mu_s(t)} \right)_{\substack{1 \leq k \leq K \\ 0 \leq t \leq K-1}} \\ &= \det \left(\begin{cases} z_k^{\mu_s(t)} - z_K^{\mu_s(t)}, & k < K \\ z_K^{\mu_s(t)}, & k = K \end{cases} \right)_{\substack{1 \leq k \leq K \\ 0 \leq t \leq K-1}} \\ &= (-1)^{K-1} \det \left(z_k^{\mu_s(t)} - z_K^{\mu_s(t)} \right)_{\substack{1 \leq k \leq K-1 \\ 1 \leq t \leq K-1}} \\ &= \det \left(\sum_{\tau_t=0}^{\mu_s(t)-1} z_k^{\tau_t} z_K^{\mu_s(t)-1-\tau_t} \right)_{\substack{1 \leq k \leq K-1 \\ 1 \leq t \leq K-1}} \prod_{k=1}^{K-1} (z_K - z_k) \\ &= \det \left(\sum_{\tau_t=\mu_s(t-1)}^{\mu_s(t)-1} z_k^{\tau_t} z_K^{\mu_s(t)-1-\tau_t} \right)_{\substack{1 \leq k \leq K-1 \\ 1 \leq t \leq K-1}} \prod_{k=1}^{K-1} (z_K - z_k) \\ &= \det \left(\begin{cases} z_k^{\mu_s(t)-1}, & t \neq s \\ z_k^{s-1} z_K + z_k^s, & t = s \end{cases} \right)_{\substack{1 \leq k \leq K-1 \\ 1 \leq t \leq K-1}} \prod_{k=1}^{K-1} (z_K - z_k) \end{aligned}$$

$$= (z_K \Delta_{K-1,s} + \Delta_{K-1,s-1}) \prod_{k=1}^{K-1} (z_K - z_k). \quad (31)$$

Similar identities hold if $s = 0$ or $s = K$, but with one of the terms $\Delta_{K-1,s-1}$, $z_K \Delta_{K-1,s}$ missing:

$$\Delta_{K,0} = z_K \Delta_{K-1,0} \prod_{k=1}^{K-1} (z_K - z_k), \quad \Delta_{K,K} = \Delta_{K-1,K-1} \prod_{k=1}^{K-1} (z_K - z_k). \quad (32)$$

Iterating identities (31),(32) K times, we get

$$\Delta_{K,s} = \prod_{1 \leq k < l \leq K} (z_l - z_k) \sum_{1 \leq k_1 < \dots < k_{K-s} \leq K} \prod_{t=1}^{K-s} z_{k_t}.$$

Substituting this equality in (30) and combining with (28) and (29), we get (27) with $q = K - s$. \square

To derive now the lower bound in (14), fix a $t_0 = t_0(K) > 0$, to be chosen later. Using monotonicity of $\hat{G}(t)$ for $t \geq 0$, which follows from (9), we write:

$$\begin{aligned} \sigma_{x;\{x_k\}_{k=1}^K}^2 &= \min_{\lambda_1, \dots, \lambda_K} \int_{\mathbb{R}} \left| e^{ixt} - \sum_{k=1}^K \lambda_k e^{ix_k t} \right|^2 \hat{G}(t) dt \\ &\geq \hat{G}\left(\frac{(K+1)t_0}{2}\right) \min_{\lambda_1, \dots, \lambda_K} \int_{-\frac{(K+1)t_0}{2}}^{\frac{(K+1)t_0}{2}} \left| e^{ixt} - \sum_{k=1}^K \lambda_k e^{ix_k t} \right|^2 dt \\ &= \hat{G}\left(\frac{(K+1)t_0}{2}\right) \min_{\lambda_1, \dots, \lambda_K} \int_{-\frac{(K+1)t_0}{2}}^{\frac{(K+1)t_0}{2}} \sum_{l=0}^K \left| e^{ixt} e^{ilxt_0} - \sum_{k=1}^K \lambda_k e^{ix_k t} e^{ilx_k t_0} \right|^2 dt \\ &\geq \hat{G}\left(\frac{(K+1)t_0}{2}\right) \int_{-\frac{(K+1)t_0}{2}}^{\frac{(K+1)t_0}{2}} \min_{\lambda_1, \dots, \lambda_K} \sum_{l=0}^K \left| e^{ilxt_0} - \sum_{k=1}^K \lambda_k e^{i(x_k-x)t} e^{ilx_k t_0} \right|^2 dt \\ &= \hat{G}\left(\frac{(K+1)t_0}{2}\right) t_0 \min_{\lambda_1, \dots, \lambda_K} \sum_{l=0}^K \left| e^{ilxt_0} - \sum_{k=1}^K \lambda_k e^{ilx_k t_0} \right|^2 \\ &= \hat{G}\left(\frac{(K+1)t_0}{2}\right) t_0 \rho^2, \end{aligned}$$

where ρ is the distance defined as in Lemma 2 for $z = e^{ixt_0}$, $z_k = e^{ix_k t_0}$. Since $|z| = |z_k| = 1$, the denominator in (27) is bounded from above by 2^K . Therefore,

$$\sigma_{x;\{x_k\}_{k=1}^K}^2 \geq \hat{G}\left(\frac{(K+1)t_0}{2}\right) \frac{t_0}{2^{2K}} \prod_{k=1}^K |e^{ixt_0} - e^{ix_k t_0}|^2.$$

Let us assume that

$$t_0 < \frac{\pi}{2}. \quad (33)$$

In this case, since $x, x_k \in [-1, 1]$, we have $\frac{|x-x_k|t_0}{2} < \frac{\pi}{2}$, hence

$$|e^{ixt_0} - e^{ix_k t_0}| = 2 \sin \frac{|x-x_k|t_0}{2} \geq \frac{4}{\pi} \frac{|x-x_k|t_0}{2} = \frac{2|t_0|}{\pi} |x-x_k|.$$

Therefore, assuming (33),

$$\sigma_{x;\{x_k\}_{k=1}^K}^2 \geq \hat{G} \left(\frac{(K+1)t_0}{2} \right) \frac{t_0^{2K+1}}{\pi^{2K}} \prod_{k=1}^K |x-x_k|^2. \quad (34)$$

Now let us choose t_0 so that

$$\frac{(K+1)t_0}{2} = e^{s_K^*},$$

where s_K^* is given by (24). Then, if (33) holds, we get from (34)

$$\begin{aligned} \sigma_{x;\{x_k\}_{k=1}^K}^2 &\geq e^{-T(s_K^*)} e^{(2K+1)s_K^*} \frac{2^{2K+1}}{(K+1)^{2K+1} \pi^{2K}} \prod_{k=1}^K |x-x_k|^2 \\ &= e^{T^*(2K+1) - (2K+1) \ln(K+1)} \frac{2^{2K+1}}{\pi^{2K}} \prod_{k=1}^K |x-x_k|^2. \end{aligned}$$

This implies the lower bound in (14), since $\frac{4}{\pi^2} > \frac{1}{e}$. We have to check, however, that condition (33) is fulfilled. The value s_K^* satisfies the condition $2K+1 = T'(s_K^*) = e^{s_K^*} S'(e^{s_K^*})$. Since $S'(t) \rightarrow +\infty$ as $t \rightarrow +\infty$, it follows that

$$e^{s_K^*} = o(2K+1) \quad \text{as } K \rightarrow \infty. \quad (35)$$

Therefore $t_0 \rightarrow 0$ as $K \rightarrow \infty$, so (33) is fulfilled for sufficiently large K .

We now prove (15). Since $T(s_K^*) \geq 0$ for sufficiently large K , we have

$$\frac{F(K)}{2K+1} = \frac{T^*(2K+1) - (2K+1) \ln K}{2K+1} = s_K^* - \frac{T(s_K^*)}{2K+1} - \ln K \leq s_K^* - \ln K \xrightarrow{K \rightarrow \infty} -\infty,$$

where we used (35) in the last step.

It remains to prove that $F(K)$ monotonically decreases for sufficiently large K . We want to show that

$$\frac{dF(K)}{dK} = 2(T^*)'(2K+1) - 2 \ln K - \frac{2K+1}{K} < 0.$$

It suffices to show that

$$(T^*)'(2K+1) - \ln(2K+1) \rightarrow -\infty, \quad \text{as } K \rightarrow +\infty.$$

By duality of the Legendre transform, this is equivalent to

$$s - \ln(T'(s)) \rightarrow -\infty, \quad \text{as } s \rightarrow +\infty,$$

which follows from (13),(10).

6 Proof of Theorem 3

In this section, $\{x_k\}_{k=1}^\infty$ denotes the optimization trajectory obtained by (2),(3) with $x_1 = 0$.

We start proving Theorem 3 by first noting that, under the hypotheses of the theorem, the mean expected value of the objective function $f = -G$ at each point of $[-1, 1]$ is exactly equal to its actual value, throughout the whole optimization process:

$$m_{x;\{x_k, f(x_k)\}_{k=1}^K} = f(x), \quad \forall x \in [-1, 1], \forall K \geq 1.$$

Indeed, by (5), $m_{x;\{x_k, f(x_k)\}_{k=1}^K}$ is the unique interpolant of the function f at the points x_1, \dots, x_K having the form $\sum_{k=1}^K \lambda_k G(x - x_k)$ with some coefficients λ_k . But $f(x) = -G(x - x_1)$ is of this form, so it is equal to the interpolant.

Since f attains its minimum at $x_1 = 0$, we have $f_K^* = f^* = -G(0)$ for all K , hence the expected improvement can be written as

$$\begin{aligned} I_{K;\{x_k, f(x_k)\}_{k=1}^K}(x) &= \mathbb{E}\left(-G(0) - \min(-G(0), \xi_x) \mid \{\xi_{x_k} = f(x_k)\}_{k=1}^K\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_{x;\{x_k\}_{k=1}^K}} \int_{-G(0)}^{-G(x)} \exp\left\{-\frac{(t + G(x))^2}{2\sigma_{x;\{x_k\}_{k=1}^K}^2}\right\} (-G(0) - t) dt \\ &= \frac{\sigma_{x;\{x_k\}_{k=1}^K}}{\sqrt{2\pi}} \int_0^\infty \exp\left\{-\frac{1}{2}\left(w + \frac{G(0) - G(x)}{\sigma_{x;\{x_k\}_{k=1}^K}}\right)^2\right\} w dw \end{aligned} \quad (36)$$

Lemma 3. For any $h \geq 0$

$$\frac{1}{2}e^{-h^2} \leq \int_0^\infty e^{-\frac{(w+h)^2}{2}} w dw \leq e^{-\frac{h^2}{2}}.$$

Proof. On the one hand,

$$\int_0^\infty e^{-\frac{(w+h)^2}{2}} w dw \leq \int_0^\infty e^{-\frac{w^2}{2}} e^{-\frac{h^2}{2}} w dw = e^{-\frac{h^2}{2}},$$

where we have used $hw \geq 0$. On the other hand,

$$\int_0^\infty e^{-\frac{(w+h)^2}{2}} w dw \geq \int_0^\infty e^{-w^2} e^{-h^2} w dw = \frac{1}{2}e^{-h^2},$$

where we have used $hw \leq \frac{w^2+h^2}{2}$. □

In the sequel, we shorten the notation for the expected improvement to $I_K(x)$.

Lemma 4. Under the assumptions of Theorem 3, there exist constants $c_1, c_2 > 0$, depending only on the kernel G , such that the following assertions hold for K large enough.

1. For all $x \in [-1, 1]$

$$I_K(x) \geq \exp\left\{-c_1 e^{K-F(K)} \frac{x^2}{\prod_{k=2}^K |x_k - x|^2}\right\} \frac{e^{(F(K)-K)/2}}{2\sqrt{2\pi}} |x| \prod_{k=2}^K |x_k - x| \quad (37)$$

$$I_K(x) \leq \exp\left\{-c_2 e^{-2K-F(K)} \frac{x^2}{\prod_{k=2}^K |x_k - x|^2}\right\} \frac{e^{F(K)/2+K}}{\sqrt{2\pi}} |x| \prod_{k=2}^K |x_k - x| \quad (38)$$

2. If, additionally,

$$|x| < \frac{1}{2} \min_{k=2,\dots,K} |x_k|, \quad (39)$$

then

$$I_K(x) \geq \exp \left\{ -c_1 (4e)^K e^{-F(K)} \frac{x^2}{\prod_{k=2}^K |x_k|^2} \right\} \frac{e^{F(K)/2} (4e)^{-K/2}}{\sqrt{2\pi}} |x| \prod_{k=2}^K |x_k| \quad (40)$$

$$I_K(x) \leq \exp \left\{ -c_2 \left(\frac{3e}{2} \right)^{-2K} e^{-F(K)} \frac{x^2}{\prod_{k=2}^K |x_k|^2} \right\} \frac{e^{F(K)/2} \left(\frac{3e}{2} \right)^K}{2\sqrt{2\pi}} |x| \prod_{k=2}^K |x_k| \quad (41)$$

Proof.

1. Since G is strictly positive definite, we have $G(0) > G(x)$ for all $x \neq 0$, and hence there exist constants $c'_1, c'_2 > 0$ such that

$$c'_1 x^2 \leq G(0) - G(x) \leq c'_2 x^2, \quad \text{for all } x \in [-1, 1].$$

Using Theorem 2, identity (36) and Lemma 3, we then get (37),(38) with $c_1 = (c'_2)^2, c_2 = (c'_1)^2/2$. Note that the $k = 1$ factor is not present in the products over k in the exponentials, as it equals x^2 and has been cancelled with x^2 in the numerator.

2. From the inequalities

$$\frac{1}{2}|x_k| \leq |x_k| - |x| \leq |x_k - x| \leq |x_k| + |x| \leq \frac{3}{2}|x_k|$$

we obtain

$$\left(\frac{1}{2} \right)^{K-1} \prod_{k=2}^K |x_k| \leq \prod_{k=2}^K |x_k - x| \leq \left(\frac{3}{2} \right)^{K-1} \prod_{k=2}^K |x_k|$$

and then substitute these latter inequalities in (37), (38).

□

Lemma 5. *Under the assumptions of Theorem 3, for all sufficiently large K :*

a)

$$I_K(x_{K+1}) \geq e^{2F(K)} \prod_{k=2}^K |x_k|^2, \quad (42)$$

b)

$$|x_{K+1}| \geq e^{2F(K)} \prod_{k=2}^K |x_k|, \quad (43)$$

c)

$$|x_{K+1}| \geq e^{2^K F(K)}, \quad (44)$$

$$|I_K(x_{K+1})| \geq e^{2^K F(K)}, \quad (45)$$

d)

$$|x_{K+1}| \leq e^{F(K)/3}. \quad (46)$$

Proof.

- a) Given K , let us choose $x \in [-1, 1]$ so as to make the expression in braces in (40) equal to -1, i.e.,

$$|x| = c_1^{-1/2} (4e)^{-K/2} e^{F(K)/2} \prod_{k=2}^K |x_k|.$$

By Theorem 2, $F(K)/K \rightarrow -\infty$, so condition (39) holds if K is large enough, and we can apply inequality (40):

$$I_K(x) \geq c_3 (4e)^{-K} e^{F(K)} \prod_{k=2}^K |x_k|^2, \quad (47)$$

with some constant c_3 depending on G . By definition of x_{K+1} , $I_K(x_{K+1}) \geq I_K(x)$. Finally, using again $F(K)/K \rightarrow -\infty$, we arrive at (42).

- b) Suppose that (43) is violated for infinitely many $K \in \mathbb{N}$. Then, for sufficiently large such K , the value x_{K+1} satisfies condition (39) with $x = x_{K+1}$, and we can apply bound (41). It follows that for such K

$$I_K(x_{K+1}) \leq \frac{e^{5F(K)/2} \left(\frac{3e}{2}\right)^K}{2\sqrt{2\pi}} \prod_{k=2}^K |x_k|^2 = o(I_K(x_{K+1})),$$

where in the last equality we have used (42) and that $F(K)/K \rightarrow -\infty$. Therefore the hypothesis that (43) is violated for infinitely many K is false.

- c) To show (44), we continue inequality (43) by iteratively applying it to x_{k+1} with $k = K, K-1, \dots, K_0+1$, where K_0+1 is the lowest value for which it is valid:

$$\begin{aligned} |x_{K+1}| &\geq e^{2F(K)} \prod_{k=2}^K |x_k| \\ &\geq e^{2F(K)+2F(K-1)} \prod_{k=2}^{K-1} |x_k|^2 \\ &\geq e^{2F(K)+2F(K-1)+4F(K-2)} \prod_{k=2}^{K-2} |x_k|^4 \end{aligned}$$

$$\begin{aligned}
& \dots \\
& \geq e^{2F(K) + \sum_{k=K_0}^{K-1} 2^{K-k} F(k)} \prod_{k=2}^{K_0} |x_k|^{2^{K-K_0}} \\
& \geq e^{2^{K-K_0+1} F(K)} \left(\prod_{k=2}^{K_0} |x_k|^{2^{-K_0}} \right)^{2^K} \\
& = e^{2^K F(K)} \exp \left\{ 2^K \left[(2^{-K_0+1} - 1) F(K) + \ln \left(\prod_{k=2}^{K_0} |x_k|^{2^{-K_0}} \right) \right] \right\},
\end{aligned}$$

where we assumed without loss of generality that $K_0 \geq 2$ and that monotonicity of $F(k)$ established in Theorem 2 holds for $k \geq K_0$. The second exponential factor in the last expression is greater than 1 for sufficiently large K due to $F(K) \rightarrow -\infty$, which implies (44).

Inequality (45) is proved in the same way, using (42) instead of (43) in the first step.

- d) Suppose that (46) is violated for infinitely many $K \in \mathbb{N}$. First, observe that for sufficiently large such K the bound (38), when applied to $x = x_{K+1}$, implies

$$I_K(x_{K+1}) \leq \exp\{-e^{-F(K)/4}\}. \quad (48)$$

Indeed, consider the first, exponential factor in (38). Using $|x_{K+1}| > e^{F(K)/3}$, the inequalities $|x_k - x_{K+1}| \leq 2$, and $F(K)/K \rightarrow -\infty$, we can write:

$$-c_2 e^{-2K-F(K)} \frac{x_{K+1}^2}{\prod_{k=2}^K |x_k - x_{K+1}|^2} \leq -c_2 (2e)^{-2K} e^{-F(K)/3} \leq -e^{-F(K)/4}$$

for K large enough. As for the remaining factor,

$$\frac{e^{F(K)/2+K}}{\sqrt{2\pi}} |x_{K+1}| \prod_{k=2}^K |x_k - x_{K+1}|,$$

it is bounded by 1 for large K , again due to $F(K)/K \rightarrow -\infty$. We thus conclude (48).

Now, combining (48) with (45), we see that

$$e^{2^K F(K)} \leq I_K(x_{K+1}) \leq \exp\{-e^{-F(K)/4}\}.$$

This implies $2^K(-F(K)) \geq e^{-F(K)/4}$. Since $c \leq e^{c/8}$ for sufficiently large c , we get $2^K \geq e^{-F(K)/8}$. But this inequality is violated for all K large enough, since $F(K)/K \rightarrow -\infty$. Therefore our assumption that (46) is violated for infinitely many $K \in \mathbb{N}$ was wrong.

□

Inequalities (44) and (46) form the statement (18) of Theorem 3. Since $F(K)/K \rightarrow -\infty$, from (46) we conclude $|x_K| \rightarrow 0$, which completes the proof.

Acknowledgement

The author thanks the two anonymous referees for the careful reading of the manuscript and several valuable suggestions and corrections.

References

- [1] A. D. Bull. Convergence rates of efficient global optimization algorithms (2011). *Journal of Machine Learning Research*, 12:2879-2904 (2011).
- [2] A. I. J. Forrester, A. Sóbester, and A. J. Keane. *Engineering design via surrogate modelling: a practical guide*. J. Wiley (2008).
- [3] H.-M. Gutmann. A radial basis function method for global optimization. *J. of Global Optimization*, 19:201–227 (2001).
- [4] D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *J. of Global Optimization*, 21:345–383 (2001).
- [5] D. R. Jones, M. Schonlau, and W. J. Welch. A data analytic approach to Bayesian global optimization. In *Proceedings of the ASA, Section on Physical and Engineering Sciences*, 186 – 191 (1997).
- [6] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *J. of Global Optimization*, 13:455–492, (1998).
- [7] Y. Katznelson. *An introduction to harmonic analysis*. Cambridge mathematical library. Cambridge University Press (2004).
- [8] D. Kincaid and E. Cheney. *Numerical analysis: mathematics of scientific computing*. Pure and applied undergraduate texts. American Mathematical Society (2002).
- [9] M. Locatelli. Bayesian algorithms for one-dimensional global optimization. *J. of Global Optimization*, 10:57–76 (1997).
- [10] J. Mockus. *Bayesian approach to global optimization: theory and applications*. Mathematics and its applications: Soviet series. Kluwer Academic (1989).
- [11] J. Mockus, V. Tiesis, and A. Zilinskas. The application of bayesian methods for seeking the extremum. *Towards Global Optimization*, 2:117 – 129 (1978).
- [12] MPMATH library: <http://code.google.com/p/mpmath/>
- [13] C. E. Rasmussen and C. K. I. Williams *Gaussian Processes for Machine Learning*. The MIT Press (2006).

- [14] M. Schonlau. *Computer experiments and global optimization*. PhD thesis, Waterloo, Ont., Canada (1997).
- [15] M. Schonlau and W. J. Welch. Global optimization with nonparametric function fitting. *Proceedings of the ASA, Section on Physical and Engineering Sciences*, 183 – 186 (1996).
- [16] A. Torn and A. Zilinskas. *Global optimization*. Springer-Verlag New York (1989).
- [17] E. Vazquez and J. Bect. Pointwise consistency of the kriging predictor with known mean and covariance functions. In *mODa 9 – Advances in Model-Oriented Design and Analysis*. 14th-19th June 2010, Bertinoro, Italy.
- [18] E. Vazquez and J. Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and Inference*, 140(11):3088 – 3095 (2010).